5

AL-TR-1991-0127

ARMSTRONG LABORATORY

# MEASURES OF SITUATION AWARENESS: AN EXPERIMENTAL EVALUATION (U)

*Martin L. Fracker, Major,* USAF

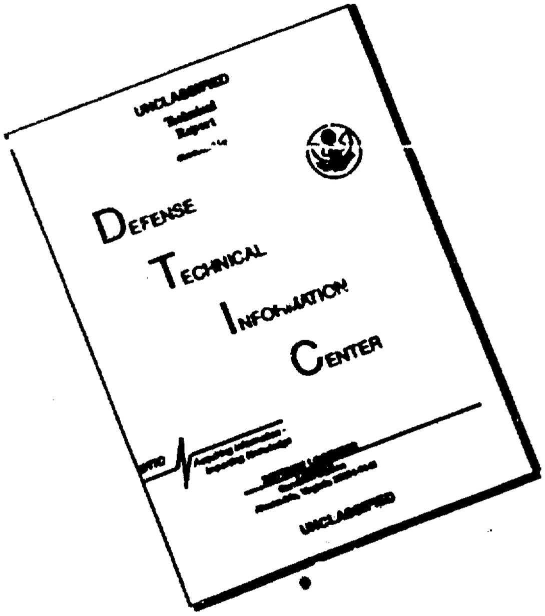
CREW SYSTEMS DIRECTORATE
HUMAN ENGINEERING DIVISION

OCTOBER 1991

FINAL REPORT FOR PERIOD JANUARY 1990 TO JANUARY 1991

Approved for public release; distribution is unlimited.

AIR FORCE SYSTEMS COMMAND
WRIGHT-PATTERSON AIR FORCE BASE, OHIO 45433-6573



## NOTICES

Please do not request copies of this report from the Armstrong Aerospace Medical Research Laboratory. Additional copies may be purchased from:

National Technical Information Service
5285 Royal Road
Springfield, Virginia 22161

Federal Government agencies and their contractors registered with the Defense Technical Information Center should direct requests for copies of this report to:

Defense Technical Information Center
Cameron Station
Alexandria, Virginia 22314

## TECHNICAL REVIEW AND APPROVAL

AL-TR-1991-0128

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

The voluntary informed consent of the subjects used in this research was obtained as required by Air Force Regulation 169-3.

This technical report has been reviewed and is approved for publication.

**FOR THE COMMANDER**

Kenneth R Boff

**KENNETH R. BOFF**, Chief
Human Engineering Division
Armstrong Laboratory

## REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

[illegible]

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
October 1991

3. REPORT TYPE AND DATES COVERED
Final Report January 1990-January 1991

4. TITLE AND SUBTITLE
Measures of Situation Awareness: An Experimental Evaluation

5. FUNDING NUMBERS
C - F33615-[illegible]
PE - 62202F
PR - 71[illegible]
TA - 1[illegible]
WU - 25

6. AUTHOR(S)
Martin L. Fracker, Major

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Human Engineering Division
Armstrong Laboratory
AL/CFHW
Wright-Patterson AFB OH 45433-6573

8. PERFORMING ORGANIZATION REPORT NUMBER
AL-TR-1991-[illegible]

9. SPONSORING MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Human Engineering Division
AL/CFHW
Wright-Patterson AFB OH 45433-6573

10. SPONSORING MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Both explicit and implicit measures of situation awareness (SA) were evaluated in a series of experiments in order to assess their reliability and two kinds of validity: criterion and construct. In all of the experiments, subjects performed a simulated combat task in which they had to monitor the positions of enemy, friendly, and neutral objects. In addition, subjects had to attack and defend themselves against enemy objects. A memory probe procedure was used to explicitly assess two components of SA: location and identity awareness. In addition, a signal detection analysis was used to provide an implicit measure of SA. Test-retest correlations indicated that the location awareness measure was much less reliable than the others. Criterion validity was evaluated by correlating the SA measures with probability of a kill in the combat task. Although the SA metrics seemed to be fairly good predictors of kill probability, the best predictor was a measure of behavioral workload. Predictions of multiple resource theory were used to evaluate construct validity. In particular, it was predicted that difficulty in maintaining identity awareness would not affect location awareness, and this prediction was largely supported. Nevertheless, other aspects of the data seemed to contradict current versions of multiple resource theory.

14. SUBJECT TERMS
mental workload　　attention
signal detection theory
memory probes　　situation awareness

15. NUMBER OF PAGES
44

16. PRICE CODE

| | | | |
|---|---|---|---|
| 17. SECURITY CLASSIFICATION OF REPORT | 18. SECURITY CLASSIFICATION OF THIS PAGE | 19. SECURITY CLASSIFICATION OF ABSTRACT | 20. LIMITATION OF ABSTRACT |
| UNCLASSIFIED | UNCLASSIFIED | UNCLASSIFIED | UNLIMITED |

# SUMMARY

Both explicit and implicit measures of situation awareness (SA) were evaluated in a series of experiments in order to assess their reliability and two kinds of validity: criterion and construct. In all of the experiments, subjects performed a simulated combat task in which they had to monitor the positions of enemy, friendly, and neutral objects. In addition, subjects had to attack and defend themselves against enemy objects. A memory probe procedure was used to explicitly assess two components of SA: location and identity awareness. In addition, a signal detection analysis was used to provide an implicit measure of SA. Test-retest correlations indicated that the location awareness measure was much less reliable that the others. Criterion validity was evaluated by correlating the SA measures with probability of a kill in the combat task. Although the SA metrics seemed to be fairly good predictors of kill probability, the best predictor was a measure of behavioral workload. Predictions of multiple resource theory were used to evaluate construct validity. In particular, it was predicted that difficulty in maintaining identity awareness would not affect location awareness, and this prediction was largely supported. Nevertheless, other aspects of the data seemed to contradict current versions of multiple resource theory.

# PREFACE

The author thanks Mike Vidulich and Mike Venturino for helpful discussions which contributed to the development of the present experiments. Of course, the author alone is responsible for any errors contained in this research. Thanks also to Sharon Davis, Ray Marcello, and Mike Ewry for their help in data collection. Finally, Brian Porter's computer programming assistance was invaluable and is gratefully acknowledged.

## TABLE OF CONTENTS

Section Page

INTRODUCTION .......... 1
Explicit and Implicit Measures of SA .......... 1
Explicit Measures .......... 1
Implicit Measures .......... 2
Evaluation of Explicit and Implicit SA Metrics .......... 3
Reliability .......... 3
Validity .......... 5
The Present Experiments .......... 6
METHOD .......... 9
All Experiments .......... 9
Experiment 1 .......... 9
Experiment 2 .......... 13
Experiment 3 .......... 15
RESULTS .......... 16
Reliability .......... 16
Criterion Validity .......... 17
Construct Validity .......... 18
SA Metrics and Avoidance Failures .......... 18
Correlations Among SA Metrics .......... 19
Experiment 1 Results .......... 20
Experiment 2 Results .......... 23
Experiment 3 Results .......... 27
DISCUSSION .......... 29
Reliability .......... 29
Criterion Validity .......... 31
Construct Validity .......... 32
Conclusions .......... 34
REFERENCES .......... 35

## LIST OF TABLES

Table Page

1. Between-session test-retest correlations.................. 17
2. Correlations between SA metrics and kill probability....... 18
3. Correlations between SA metrics and avoidance failures..... 19
4. Theoretical upper limits of correlations among SA measures. 20
5. Obtained correlations among SA measures................... 20
6. Combat task data from Experiment 1........................ 21
7. Location error from Experiment 1.......................... 22
8. Identity probe data from Experiment 1..................... 23
9. Combat task data from Experiment 2........................ 24
10. Location error from Experiment 2.......................... 25
11. Identity probe data from Experiment 2..................... 26
12. Memory task data from Experiment 2........................ 27
13. Combat task data from Experiment 3........................ 28
14. Location error from Experiment 3.......................... 28
15. Identity probe data from Experiment 3..................... 28
16. Memory task data Experiment 3............................. 29

## INTRODUCTION

Situation awareness (SA) refers to military operators' knowledge of their immediate tactical situation. Clausewitz (1832/1984) pointed to the importance of SA when he wrote that the "difficulty of *accurate recognition* constitutes one of the most serious sources of friction in war, by making things appear entirely different from what one had expected" (p. 117). Poor awareness of the tactical situation generally leads to surprise and, as Clausewitz observed, surprise is one of the principal means by which one side or the other gains the advantage in war (p. 198). Thus, maintaining ones own SA while depriving the enemy of his is a necessary condition for victory in war.

Because of the centrality of SA to combat operations, the Air Force has invested considerable effort in improving operator SA through improved command, control, and communication systems, better cockpit displays, and more effective pilot training. In addition, efforts to develop "SA countermeasures"--intended to inflict poor SA on the enemy--are presumably also underway. In order to evaluate the success of these efforts, measures of operator SA are needed. The present series of experiments was intended to begin addressing this need. Two classes of measures were examined: explicit and implicit. These two classes are first described in general, and then criteria against which the measures may be evaluated are outlined. Following this general discussion, the specific issues and measures examined in the present experiments are described.

### Explicit and Implicit Measures of SA

The distinction between explicit and implicit measures comes from a distinction made by some psychologists between explicit and implicit forms of memory (see Roediger, 1990). Explicit measures require people to self-report material in memory of which they are consciously aware. As a result, such measures are considered subjective in nature but should be distinguished from subjective rating measures, which involve assignment of numerical values to the *quality* of the content of awareness. Unlike explicit measures, implicit measures do not rely on self-reports of awareness; rather, such measures are derived from task performance. Specifically, SA is inferred from the influence of prior events on task performance (e.g., evading an attacking aircraft, locking on to an enemy target). Thus, implicit measures may be considered objective rather than subjective in nature.

#### Explicit Measures

If SA is regarded as the information immediately available in conscious awareness, then explicit measures are the most direct way of assessing SA. Two types of explicit measurement methods can be identified: retrospective event recall (e.g., Kibby, 1988; Whitaker and Klein, 1988) and concurrent memory probes (Endsley, 1988, 1989). In retrospective measures, operators perform a mission first and then recall facts about the mission afterwards. Perhaps the most serious challenge to this approach is the possibility of false memories. A growing body of research shows that as the time between an event and its recall increases, people become more likely to recall "facts" about the event that in fact are not true (Loftus, 1979; Loftus and Loftus, 1980). These false recollections appear to be otherwise reasonable inferences drawn from information that people are still able to remember (Carr, 1986). Because progressively more information is forgotten as time goes on, such false inferences increase in frequency as the event becomes more distant. Thus, retrospective recall seems as likely to measure what operators can infer may have happened as what they can remember having actually happened.

To avoid problems such as false recollection, Endsley (1988, 1989) has suggested the use of what may be called concurrent memory probes. In her approach, memory is probed closer to the time specific events actually occur--during the mission rather than afterwards. Other investigators have also used memory probes to assess SA, including Marshak, Kuperman, Ramsey, and Wilson (1987), Venturino and Kunze (1989), and Wells, Venturino, and Osgood (1988). The basic idea in most of these implementations is to freeze a simulated mission after some random interval of time, blank the pilot's displays, and ask the pilots to recall certain items of information, such as the locations of particular aircraft. SA is then quantified as the pilot's error in responding to these queries.

Perhaps the major disadvantage of memory probes is their intrusiveness: one must interrupt the pilot in order to administer the probe. Endsley (1989) provided data suggesting that the consequences of this intrusion may be less than one might expect, but the fact of the intrusion should not be lightly dismissed.

### Implicit Measures

Perhaps because explicit measures have clear liabilities, some researchers have focused on developing implicit alternatives (Eubanks and Killeen, 1983; Venturino, Hamilton, and Dvorchak, 1989). In implicit measurement, the goal is to determine whether pilots' mission performance has been influenced appropriately by the occurrence of specific events. The most straightforward approach uses signal detection theory to derive an SA metric (Eubanks and Killeen, 1983).

Suppose that event X occurs. If pilots are aware of the event's occurrence, then they should respond in one way (a "hit"); but if pilots are unaware that the event occurred, then they should respond in a clearly different way ("miss"). Unfortunately, the interpretation of hits and misses is always complicated by response bias. For example, pilots may be biased to attack other aircraft when they are unsure whether the aircraft is friend or foe. In order to identify and correct for such bias, it is necessary to also measure false alarms (responding as if the event occurred when it did not) and correct rejections (not responding when the event did not occur). Once these four types of responses have been identified and counted over the course of a mission, there are several methods available for computing the pilots' ability to discriminate occurrence from non-occurrence of the target event, referred to as *sensitivity* (Macmillan and Creelman, 1990). Because sensitivity declines if pilots are unaware of events occurring and increases if they are so aware, the measure provides both an empirical and an intuitively reasonable measure of awareness for a particular kind of event (cf., Hawkins et al., 1990).

Any discrete measure of performance can be used to measure sensitivity providing that target events as well as the responses to be counted as hits are unambiguously defined; that is, the presence and absence of both are clear and countable. In meeting this condition, the main challenge may often be to find response measures that react to the events of interest. Fortunately, for some kinds of events, appropriate measures are not hard to find. Eubanks and Killeen (1983) were interested in whether novice F-4E pilots would detect the entry of enemy targets into their weapon envelope. Hits, misses, false alarms, and correct rejections were defined in terms of whether or not there was an enemy in the envelope, and whether or not pilots fired the weapon.

## Evaluation of Explicit and Implicit SA Metrics

The two principle criteria by which SA metrics should be evaluated are their reliability and validity. Additional criteria such as ease of use and operator acceptance should be considered only when choosing between two or more metrics that are approximately equally reliable and valid. *Reliability* concerns whether a metric will remain consistent if the same quantity is measured at different times under the same conditions. *Validity* mainly concerns whether the metric actually measures what it is supposed to measure. Both are important. On one hand, the validity of a measure is limited by its reliability. On the other, there is nothing to prevent a highly reliable metric from being invalid. For example, measuring the length of pilots' feet is likely to provide highly reliable but completely invalid assessments of their skill in combat.

### Reliability

Reliability theory revolves around the concept of a *true score*, defined as the outcome of all the factors that influence the attribute being measured. Concerning SA, these factors might include characteristics of human operators such as their natural intelligence, training, and experience, as well as characteristics of the environment such as the availability and formatting of relevant information. Any given measure, *X*, of the attribute is then said to be the sum of the true score, *T*, and some random error in the measurement, e. The variability of *X*, then, is the variability of the sum ($T + e$). Assuming that *T* and *e* are uncorrelated, this variability can be re-expressed as the sum of Var($T$) and Var(e), denoting the variabilities of *T* and *e*, respectively. The reliability, or consistency, of a measure may be defined as the following ratio: Var($T$) / [Var($T$) + Var(e)]. Reliability improves as variability due to measurement error declines. Conversely, any factor that increases measurement error reduces reliability (for extended discussions, see Allen and Yen, 1982; Gulliksen, 1950; Lord and Novick, 1968; Murphy and Davidshofer, 1991).

Reliability can be estimated using test-retest, alternate forms, split-half, and internal consistency methods. Test-retest methods require collecting the same measure from the same people under the same conditions at different times. Assuming that the measured attribute does not change over time and that the first measurement does not influence the second, the correlation between the two measurements is a direct estimate of the measure's reliability. In alternate forms methods, two alternate versions of the same measurement technique are used on the same people and compared. Reliability is then estimated by the correlation between the two versions. Split-half methods are appropriate when a measure is aggregated from several response samples, referred to as items. Essentially, the set of items are divided in half and the correlation between the two halves is determined. Internal consistency methods estimate reliability from the intercorrelations among all of the items contributing to a measure.

The easiest way to improve the reliability of a measure is to increase the number of observations that contribute to the measure. If the observations are added together to form a composite score, then the sum will be at least as reliable as the least reliable observation. Further, if the observations are correlated, then the reliability of their sum will increase (1) as the number of observations increases and (2) as the correlations among observations are strengthened. Thus, a good way to improve the reliability of a measure is to obtain a larger number of correlated observations and use their sum (or average) as the measure.

In contrast to composite scores (sums or averages), profile scores decrease in reliability as the correlations among observations increase. P ofile scores are measures of how one variable differs from another. For example, one might measure pilots' awareness of the locations of enemy aircraft, enemy surface-to-air missiles, and enemy tanks. Some pilots might have good awareness for aircraft locations but poor awareness for missiles and tanks. Other pilots might have poor awareness for aircraft but good awareness for missiles and tanks.

Thus, looking at SA profiles might reveal specific weaknesses in SA for specific pilots. Comparing profiles is essentially equivalent to comparing differences between variables (e.g., SA for aircraft versus SA for tanks). If two variables are correlated, then they tend to reflect the same true score. Thus, subtracting one from the other will tend to leave only the random error. As a result, differences between correlated variables will tend to be highly unreliable. In general, then, profile scores should be avoided. When possible, composite scores should be used instead.

**Validity**

Validity is not a simple concept. At least three types of validity may be identified: content, criterion, and construct.

*Content validity.* Content validity refers to the degree that the knowledge or behaviors assessed by a metric represent the knowledge or task domain being measured. Assessing content validity usually involves analyzing the specific knowledge or behaviors relevant to the domain and rendering a judgment as to whether the sampled knowledge or behaviors are in fact representative. In SA measurement, establishing content validity first requires analyzing a given military task in order to determine what kinds of information the operator needs to know. This information, once determined, can then be compared to the information sampled by the SA metric. Content validity would be considered high if all important kinds of information in the domain--and no irrelevant domains of information--are sampled by the metric.

*Criterion validity.* Criterion validity refers to the degree of correlation between the metric and some objective measure that could be used to evaluate the accuracy of a decision based upon the metric. For example, if the SA metric is to be used to select one of several competing cockpit designs for a new fighter aircraft, the criterion might be success in combat.

Establishing criterion validity is usually complicated by the fact that many factors may contribute to the criterior measure. Combat success, for instance, depends not only upon accurate SA but also upon wise decision making and effective response execution. While wise decisions and effective responses are dependent upon accurate SA, possessing the latter is no guarantee that the others will follow. Thus, an otherwise valid measure of SA might appear poor if it is tested on operators who make poor decisions or unskilled responses. This observation implies a dilemma in establishing the criterion validity of SA metrics. If inexperienced or only partially trained operators are included in the study, the correlation between measured SA and the cri erion may appear low for reasons that have nothing to do with the SA metric itself. On the other hand, if only experienced and highly trained operators are included, a high correlation may be precluded for purely statistical reasons (restriction of range). Paradoxically, then, criterion validity--which is often the most important form of validity to the user--may be the most difficult to establish and hence the least likely to be assessed.

*Construct validity.* A construct is some unobservable psychological attribute such as situation awareness that is hypothesized to account for some aspect of human behavior. Construct validity refers to the degree that a measure can quantify this unobservable psychological attribute. Assessing construct validity involves identifying (1) human behaviors that are logically related to the construct in question, (2) other constructs that are either related or unrelated to the target construct, and (3) behaviors that are logically related to these new constructs (Murphy & Davidshofer, 1991). One then demonstrates that behaviors related to the construct (a) behave as they are supposed to, (b) associate with other related behaviors, and (c) dissociate from behaviors unrelated to the construct. Because statements of the relationship between specific behaviors and a given construct are theoretical in nature, tests of construct validity may also be viewed as tests of the underlying theory. Consequently, failures to establish construct validity are invariably ambiguous. Such failures may mean that the measure is invalid, or that the underlying theory is incorrect. If tests of several alternative measures within the same theoretical framework all fail to establish their construct validity, then one may conclude that the underlying theory is at least not very useful.

## The Present Experiments

The present set of experiments were intended to evaluate the reliability, criterion validity, and construct validity of memory probe (explicit) and envelope sensitivity (implicit) measures of SA. In order to conduct this evaluation, a laboratory task was devised to simulate certain aspects of combat. In this task, subjects monitored the spatial locations of several moving objects, each having a unique shape. Some of the objects were red (enemies), some were blue (friendlies), and some were gray (neutrals). In addition to monitoring object locations, subjects had control over one of the friendlies and had to use it to destroy enemies before being destroyed themselves. Subjects' performance in detecting and attacking enemies within their weapon envelope was assessed using the envelope sensitivity metric (computation is described in the Results section).

SA needed to perform the combat task successfully can be analyzed into at least two components: (1) keeping track of object locations and (2) observing each object's identity. Memory probes were used to assess subjects' SA for these two components separately. Periodically, the combat task would freeze and one of the objects would disappear from the screen. The missing object reappeared in a box at the bottom of the screen without its color. Subjects were required either to show where the missing object had been located (location probe) or to indicate its identity (identity probe).

Reliability was evaluated by having subjects perform under the same experimental conditions on two consecutive days and obtaining the correlation between sessions within each experimental condition. Thus, test-retest correlations were obtained. Criterion validity was evaluated by determining the probability of a kill in the combat task and then obtaining its correlation with each SA metric.

In order to evaluate construct validity, predictions were generated from two hypotheses. First, object locations and identities were hypothesized to be maintained in separate working memories: locations in a spatial memory and identities in a verbal memory (cf., Baddeley, 1986; Schneider and Detweiler, 1987). Second, these two working memories were hypothesized to have separate processing resources, so that increasing processing difficulty in one would have no effect on the other (Wickens, 1980, 1990; see also Friedman & Polson, 1981; Wickens & Liu, 1988; Wickens, Sandry, & Vidulich, 1983). In order to test these hypotheses, both Experiments 1 and 2 manipulated the difficulty of remembering objects' identities by causing the objects to change identities several times during a trial. As objects successively changed identities, it was expected that their previous identities would interfere with the new ones, thereby increasing the effort needed to maintain these identities in memory. This identity inconsistency was therefore expected to degrade performance on identity probes but not location probes.

A further test of the hypotheses involved manipulating the intensity of the combat task by increasing the number of enemy objects (the number of neutrals simultaneously decreased so that the total number of objects remained constant). Increasing combat intensity was expected to increase the resource demands of the combat task, affecting both spatial and verbal processes. Thus, both location and identity probes were expected to be affected. Importantly, an interaction between combat intensity and identity inconsistency was expected in probes of identity. Because increasing the intensity should overtax the available resources, fewer resources could be allocated to coping with interference caused by identity inconsistency. Thus, inconsistency was expected to have a larger detrimental impact on identity probe performance when combat intensity increased.

While the above hypotheses pertain directly to the explicit memory probe measures, they also lead indirectly to predictions regarding envelope sensitivity as well. Specifically, envelope sensitivity should be influenced by both identity and location awareness: i.e., whether the subjects' know which objects are hostile and where those objects are located. Therefore, it was predicted that envelope sensitivity would be affected by both combat intensity and identity inconsistency, showing the same interaction predicted for identity probes.

As in any test of construct validity, failures of the SA measures to behave as predicted would be ambiguous: the SA metrics might be invalid, the manipulations might not have been successful, or the underlying theory might be incorrect. To help reduce this ambiguity, a "secondary" task was embedded within the combat task. Several "danger" zones were placed throughout the combat area which subjects were instructed to avoid; failure to avoid these regions resulted in their "destruction" and a loss of points. If manipulation of combat task intensity increased resource demand, then subjects were expected to meet the increased demand by reallocating resources away from the embedded avoidance task; the frequency of avoidance failures was thus expected to increase. At the same time, because the avoidance task primarily involves spatial processing, avoidance failures should not be affected by increases in verbal processing difficulty. Therefore, identity inconsistency was not expected to affect avoidance failure frequency.

Experiments 2 and 3 contained additional tests of the hypothesis that location and identity probes tapped into separate working memories which can be characterized as spatial and verbal, respectively. One alternative is that the resources used by location and identity probes are indeed separate but are not accurately described as spatial and verbal. Another alternative is that all components of task performance and SA draw upon a common, single processing resource (cf., Kahneman, 1973; Kantowitz, 1985). In Experiment 2, subjects performed a running memory task concurrently with the combat task. For some subjects, the memory items were letters of the alphabet; for others, the items were the locations of a randomly moving cursor. Of course, the perceptual demands of attending to an additional task might interfere with other tasks regardless of whether the same central resources are used. Therefore, the variable of interest was the number of items that had to be remembered in the memory task (setsize), not simply whether the memory task had to be performed. If location and identity resources are separate and correctly characterized as spatial and verbal, then letter task setsize should affect identity but not location probe performance while spatial task setsize should affect only location probe performance.

In Experiment 3, subjects performed the two kinds of running memory tasks simultaneously in order to explicitly test the hypothesis that verbal and spatial resources are indeed separate. If they are separate, then performance on one memory task should not be influenced by setsize in the other memory task. If one memory task is effected by the other task's setsize, then it would seem likely that they share a common working memory.

As a further evaluation of construct validity, the correlations among location error, identity accuracy, and envelope sensitivity were obtained. If location error and identity accuracy measure different components of SA and are supported by independent processing resources, then they should be uncorrelated with each other. On the other hand, envelope sensitivity was expected to be correlated with both location error and identity accuracy.

## METHOD

### All Experiments

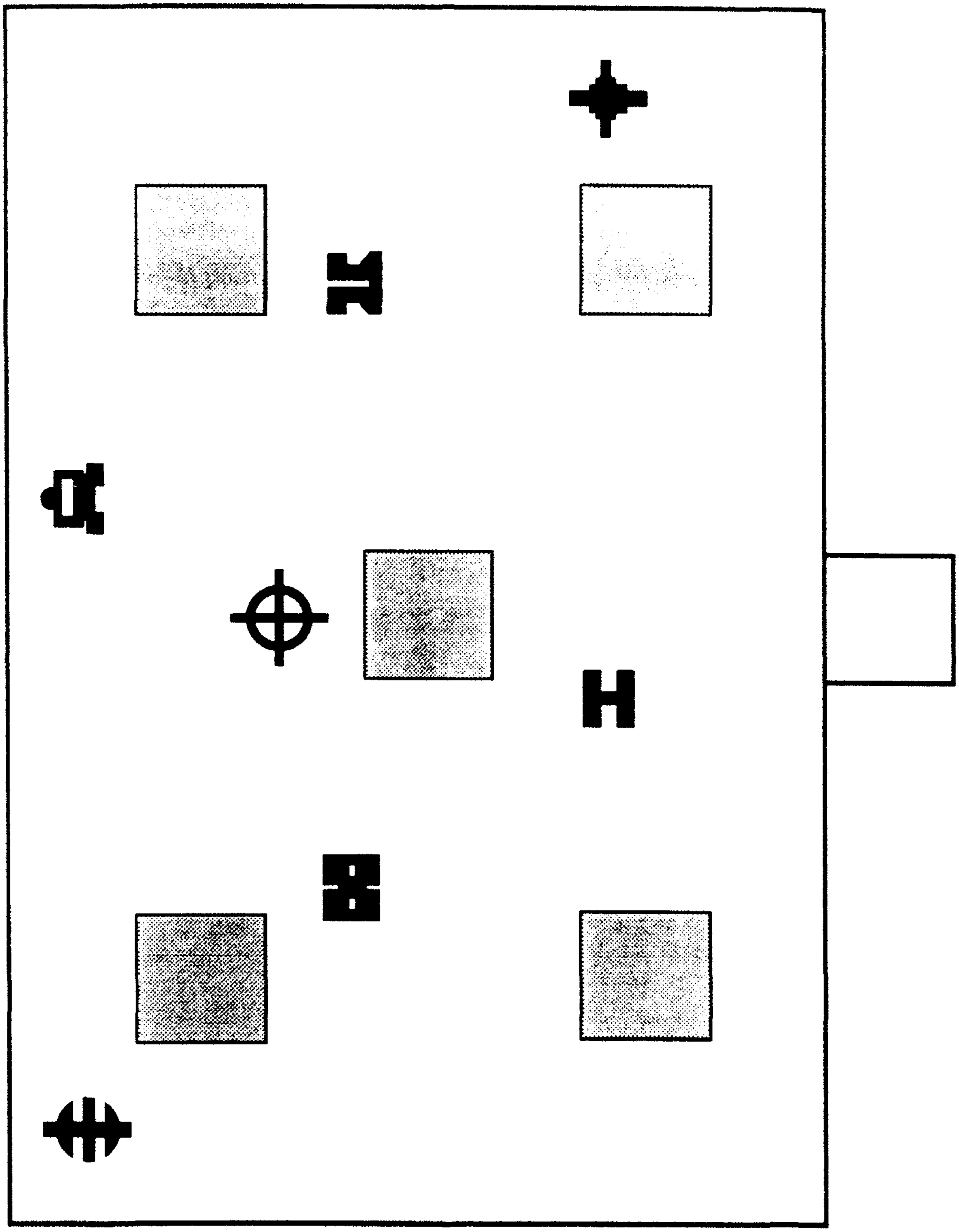
*Subjects.* In all experiments, the subjects were paid volunteers from the Wright State University community. Eight subjects participated in Experiment 1, 24 subjects participated in Experiment 2, and 9 subjects participated in Experiment 3. All subjects had normal color vision.

*Tasks.* Subjects performed the combat task in each experiment and also responded to both location and identity probes. Subjects in Experiments 2 and 3 also performed one or two additional running memory tasks. These tasks are described below.

### Experiment 1

*Combat task.* The display for the combat task is shown in Figure 1. The playing area was outlined by a large cyan colored box measuring 24.4 x 16.5 cm on a black background. Immediately below the playing area was a smaller box used for the location and identity probes; this box measured 2.5 x 2.5 cm. Distributed throughout the playing area were five dark gray 2.5 x 2.5 cm rectangular regions whose purpose is described in a later paragraph. Seven uniquely-shaped objects appeared within the playing area. The various shapes may be examined in Figure 1. All but the cross-hair shaped object were approximately 1.3 x 1.3 cm in size; the cross-hair shaped object was slightly larger, measuring 1.9 x 1.9 cm. Because subjects viewed the display from a distance of about 60 cm, the cross-haired object subtended about 1.8 degrees of visual angle; the remaining objects subtended about 1.2 degrees. Objects were colored red, blue, or light gray. Light gray objects were still clearly visible when overlaying the dark gray rectangular regions.

Subjects controlled the cross-hair shaped object. The remaining six objects were computer-controlled. One of the six objects was colored blue (as was the subject's) and was designated as "friendly." The other objects were either red, designated as "enemies," or gray, designated as "neutrals." The subject's task was to destroy red objects. This was done by using a joystick to move the cross-hair object close to a red object and pressing the joystick's fire button. Each red object had an imaginary envelope surrounding it within which



2a

it could be destroyed. This envelope extended approximately two degrees of visual angle in all directions from the center of the object. Once the subject entered into this envelope, he or she had 500 milliseconds in which to destroy the red object. If the subject failed to destroy the red object during this period, then the subject's object was destroyed. The friendly object also attacked red objects; whether it or the red object was destroyed was randomly determined. Whenever any object was destroyed, it exploded (with an appropriate sound effect) and then reappeared at some random location on the screen.

How objects moved depended upon their color. Red objects always moved in a straight line towards the nearest blue object. The direction of movement was continuously recalculated. The computer-controlled blue object always moved in a straight line towards the nearest red object. Gray objects moved about randomly. These movement strategies, coupled with the subject's requirement to attack red objects, caused red objects to be nearest to the subject's object while gray objects were generally furthest away. As a result, red objects were most likely to be within the subject's field of attention while gray objects were the least likely.

The five rectangular regions shown in Figure 1 were stationary "death zones": if the subject's object came into contact with one of these, it was destroyed and randomly relocated. Subjects were instructed to avoid these death zones. These death zones were created in order to prevent subjects from adopting a strategy of sitting still while red objects came to them. (A pilot study had shown that subjects would adopt this strategy if allowed to do so, presumably in order to make the task easier to perform.) If subjects did adopt this strategy, the computer immediately began moving their cross-hair in a straight line towards the nearest death zone.

Three measures were derived from combat task performance: kill probability, envelope sensitivity, and avoidance failures. In order to compute the first two measures, the following data were collected: the number of times subjects destroyed a red object ("hits"), the number of times subjects were themselves destroyed by red objects ("misses"), the number of times subjects fired when no red objects were close enough to be destroyed ("false alarms"), and the number of cycles in the computer program during which the fire button was not pressed *and* no red objects were close enough to be destroyed ("correct rejections"). Kill probability was defined as the the number of hits divided by the sum of hits and misses. Envelope sensitivity was computed as A' from kill probability and the false alarm rate (Pollack and Norman, 1964; see also Craig, 1979; Macmillan and Creelman, 1990). Finally, avoidance failures were measured simply as the number of times subjects ran into a death zone and were destroyed.

In addition to the preceding measures, the number of times subjects inadvertently destroyed friendly and neutral objects was also counted. Because the number of these "friendly fire accidents" was virtually zero, they will not be discussed further.

Duration of a combat task trial was 180 s plus the time needed to perform the location and identity probes. These two tasks were performed during interruptions which were not counted against the 180 s duration. The duration of the combat task between interruptions varied randomly between 10 s and 50 s with an average duration of 30 s. As a result, subjects could not predict exactly when an interruption would occur.

*Location probe.* Each combat task trial was interrupted six times, once for each of the six objects. At the moment of the interrupt, the task froze and all of the objects turned white so that their colors were no longer available to the subject. Simultaneously, one of the objects disappeared and then reappeared in the box at the bottom of the screen shown in Figure 1. Subjects used the joystick to move the object back to its correct location. Subjects moved the object in two phases: a rapid- movement phase and a precision-movement phase. During the rapid- movement phase, the joystick moved the object across the screen quickly but allowed only gross control over its location. When the subject pressed the fire button once, the precision-movement phase began. Now the joystick moved the object slowly but gave subject's control over the object's exact position on the screen. The color of the border surrounding the screen changed from one phase to the next so that subjects could tell which phase the joystick was in. Subjects were allowed unlimited time in order to position the object. When the subject pressed the firebutton a second time, the position of the tested object was recorded. Immediately, the object automatically returned to its correct location. The program then paused for 2 s so that the subject could note the object's correct location.

The subject's error in performing the location probe was calculated as the Euclidean distance in pixels between the subject's placement of the object and its correct location. Later, this distance was converted to degrees of visual angle. The recorded location errors were then averaged in two ways: first, a global average across all location probes was obtained, referred to as "average location error"; second, location errors were also averaged by identity. Thus, each combat task trial produced a global average error as well as separate average errors for friendly, enemy, and neutral objects.

*Identity probe.* Like the location probe, the identity probe was performed six times during interruptions of a combat task trial. The identity probe began in the same way as the location probe: the combat task froze, all of the objects turned white, and one of the objects disappeared and then reappeared in the box at the bottom of the screen. Simultaneously, three letters also appeared above the box area: F, H, and I, for friendly, enemy (hostile), and neutral (indifferent). Letters rather than blocks of color were used because it was thought that subjects might have encoded objects' colors by their semantic meaning as "friend," "enemy," and "neutral." The letter E was not used for enemy because it was judged too easily mistaken for F. Similarly, N was not used for neutral because it was considered confusable with H. The order in which the three letters appeared from left to right varied randomly from freeze to freeze. An arrow simultaneously appeared below the middle letter. Subjects used the joystick to move the arrow to the letter representing the object's identity and then pressed the firebutton. One of two tones then sounded to indicate to the subject whether his response was correct.

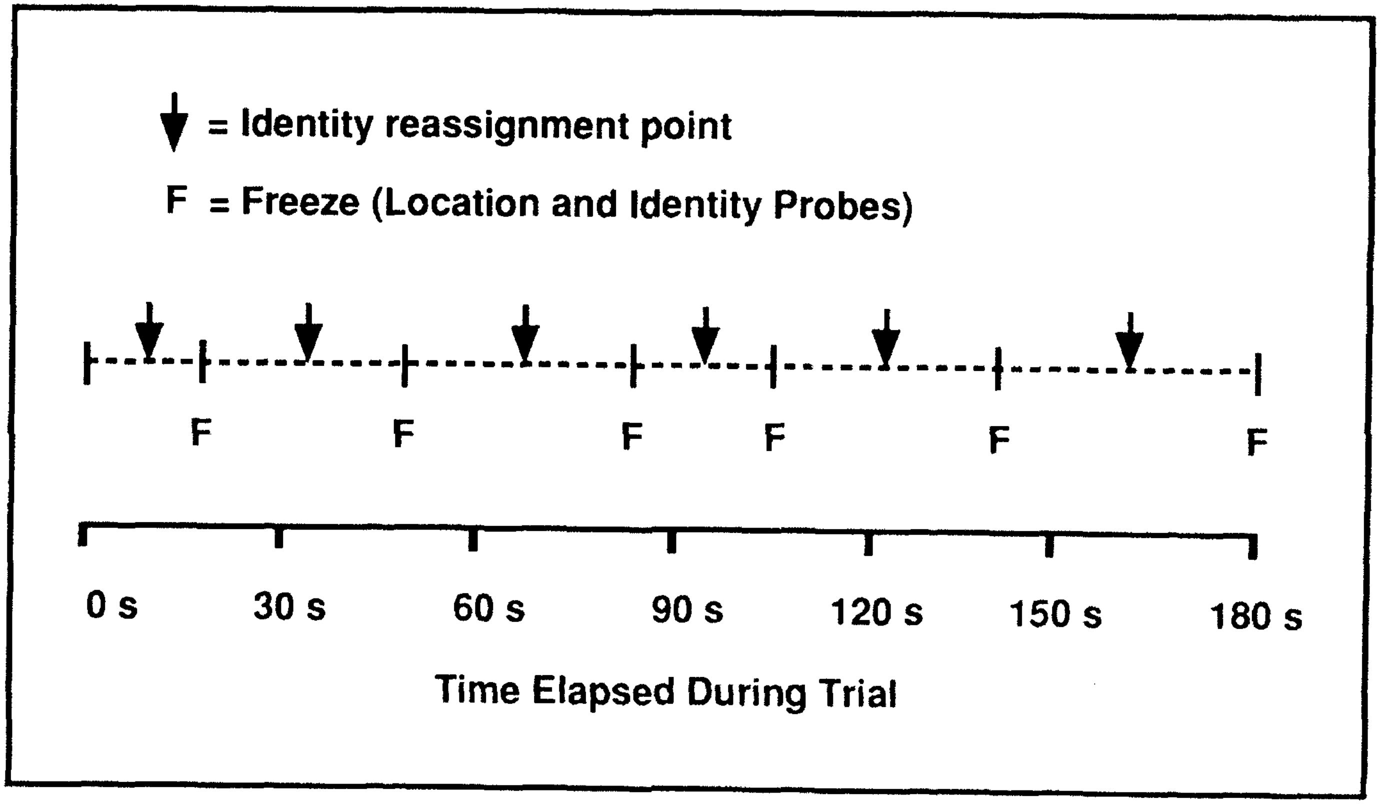
Two measures of the identity probe performance were taken. First, the subject's reaction time was recorded. Reaction time was measured from the appearance of the test display to the moment the subject first deflected the joystick or pressed the firebutton, whichever came first. This way of measuring reaction time was used both because it was convenient and because a pilot study had found this reaction time measure to be useful (it was sensitive to the number of items to be remembered in a running memory task). Second, whether the subject's response was correct or incorrect was also recorded.

*Order of location and identity probes.* During each freeze, half the subjects performed the location probe first while the other half performed the identity probe first.

*Number of enemies.* On half the trials, there was only one red object and four gray objects. On the other half, there were three red objects and only two gray objects. There were always two blue objects, one controlled by the subject and one controlled by the computer. As a result, a total of seven objects were always displayed on the screen.

*Identity consistency.* Identities were always assigned at random to the six computer-controlled objects at the beginning of every combat task trial, with the constraint that the appropriate numbers of objects be friend, enemy, and neutral. On half the trials, this assignment of identities remained intact for the full trial. On the other half, identities were randomly reassigned at six times during the trial.

The timing of identity reassignments is shown in Figure 2. These reassignments always occurred midway into the inter-freeze intervals during which subjects performed the combat task. For example, if the first freeze occurred 10 s into the combat task, the first reassignment occurred 5 s into the task. If the second freeze came 50


= Identity reassignment point
F = Freeze (Location and Identity Probes)
F
F
F
F
F
F
0 s
30 s
60 s
90 s
120 s
150 s
180 s
Time Elapsed During Trial

s after the end of the first, then the second color reassignment occurred 25 s into the second interval. There was a slight risk that this timing would cue the subject as to when the next freeze would occur, but the risk seemed outweighed by the need to control the time each object was assigned to successive colors. In any event, the average duration of a color assignment was 30 seconds. Further, at the occurrence of a freeze, an object had always been assigned its new identity for less time than it had been assigned its previous identity (see Figure 2).

*Procedure.* All subjects performed in all four experimental conditions formed by the factorial combination of the enemy number and identity consistency manipulations. Subjects were tested one at a time in each of three sessions conducted on different days. The first session was a practice day intended to familiarize subjects with the tasks and each of the four experimental conditions. During this practice session, subjects began with the easiest condition (one enemy, consistent identities) and progressed to the most difficult (three enemies, inconsistent identities). The two conditions with one enemy were performed first. Data from this session were not analyzed.

Sessions 2 and 3 were the data collection sessions. Reliability coefficients were determined by obtaining correlations between these two sessions. Subjects performed in two replications of each of the four conditions in a counterbalanced order determined an 8 x 8 Latin square.

**Experiment 2**

In addition to the task and probes of Experiment 1, subjects in Experiment 2 also performed a memory task. For half the subjects, the memory stimuli were letters of the alphabet. For the other half, the stimuli were spatial locations of a randomly moving object.

*Memory stimuli.* For the verbal memory task, the stimuli were the following nine upper case letters: B, E, G, L, R, S, U, X, and Z. For the spatial memory task, the stimulus was a large white cross-hair. The cross-hair was identical to the subject- controlled cross-hair used in the combat task except that it was (1) white rather than blue and (2) twice the size in both height and width subtending 3.6 rather than 1.8 degrees of visual angle. Because of the difference in color and size, the white cross-hair was not confusable with the subject-controlled blue cross-hair.

*Procedure.* At the beginning of combat task trial, a memory stimulus appeared on the screen. Verbal stimuli appeared in the box centered below the combat task playing area (see Figure 1). Each letter appeared for 2 s and then was replaced by the next letter. The 2 s duration was intended to give subjects enough time to orient to the letter before it disappeared. In order to alert the subject to the fact that a new letter had appeared, a short beep sounded each time a new letter was displayed. In additional to this auditory cue, two visual cues were also provided. First, successive letters alternated between positions slightly to the right and left of the center of the box. This small movement was expected to be detectable by peripheral vision enabling the subject to momentarily orient attention to the letter in order to encode and remember it. Second, the letters also alternated in color between green and yellow. These auditory and visual cues were intended to help subjects manage the visual attentional requirements of the verbal memory task.

Timing and cuing of the spatial stimuli were similar. The large white cross-hair initially appeared at a random location and remained there for 2 s. Then the cross-hair jumped to a new random location anywhere within the combat task playing area, where it remained for 2 s. Each jump of the cross-hair coincided with a short beep indicating that the cross-hair had moved. A major difference between the verbal and spatial memory tasks was that the subjects had to search for the white cross-hair (because its movement was unpredictable) whereas they did not have to search for verbal memory stimuli (which always appeared in the box). Coloring the cross-hair white against the black background of the playing area, and making the cross-hair twice the size of any other object on the screen, were intended to reduce the importance of this difference. It was expected that the large white cross-hair would easily attract attention and so be easy to find.

In both memory tasks, new stimuli continued to appear every 2 s as long as there were at least 2 s left before an interruption of the combat task. If less than 2 s remained, the last stimulus disappeared and was not replaced. Because the interval between interruptions varied randomly by 10 and 50 s in duration, the number of stimuli presented in a sequence varied from 5 to 25. Thus, subjects could safely ignore the first three stimuli in the sequence (although they were not told this). From the fourth (or fifth) stimulus on, the interruption could occur at any time so subjects had to continuously update their memory of the one or two most recent stimuli, depending on the memory load condition. As the number of stimuli in the sequence increased, interference from previous stimuli was expected to increase. This interference may explain why pilot subjects were unable to remember more than two stimuli in this task.

When an interruption occurred, subjects not only performed the location and identity probes, but were also tested on the memory task. The order of these three events within an interruption was counterbalanced across subjects within each session. A given subject always received the same order in a session, but the same subject received different orders in different sessions.

The spatial memory test procedure was similar to that used in the location probe. First, the white cross-hair appeared in the box at the bottom of the screen. Simultaneously, a message appeared to the left of the box asking the subject to recall either the last or second to last location. Subjects then used the joystick to move the crosshair to its tested location in the same manner as in the object location probe. Error was calculated as the Euclidean distance between the correct location and the subject's placement of the crosshair.

The verbal memory test was similar to the identity probe. All nine letters appeared in a row near the bottom of the playing area with an arrow below the row pointing up to the middle letter. Subjects then used the joystick to select the correct letter. Reaction time and accuracy were computed in the same way as in the identity probe.

When subjects had to remember the two most recent stimuli, subjects were instructed to recall either the last or the second to last location or letter. Whether the last or second to last stimulus was requested was randomly determined. Thus, subjects had to have both stimuli available in memory because they could not predict which one they would be asked to retrieve.

Whether the memory loading task was verbal or spatial was the only between subject variable. Thus, each subject performed in all 12 experimental conditions formed by the factorial combination of memory load (0, 1, or 2 items), enemy number (1 or 3), and identity consistency (consistent or inconsistent). The order in which subjects received these 12 conditions was counterbalanced across subjects within sessions two and three using a different Latin square for each session. Subjects received each experimental condition only once during a session.

**Experiment 3**

Experiment 3 differed from previous experiments in two ways. First, subjects performed the verbal and spatial memory tasks simultaneously. Second, enemy number and identity consistency were not manipulated; rather, there were always three enemies and identities were always inconsistent.

The verbal and spatial memory tasks were identical to those in the second experiment except that subjects sometimes performed the two together. There were nine experimental conditions formed by the factorial combinations of three load levels for each memory task: 0, 1, and 2. When both memory loads were zero, subjects performed the combat, color, and location probes alone. When the load for one memory task was zero and the other non-zero, then only one memory task was performed.

*Procedure.* Subjects again participated in three sessions. The first session was practice only. Data were collected in sessions two and three. Each subject participated in all nine experimental conditions. The order in which subjects received the nine conditions was counter-balanced using a Latin square. During the freezes, the order in which subjects responded to the location, color, and two memory task probes (if appropriate) was random.

*Apparatus.* All three experiments were controlled by a Commodore 128 computer using a Commodore 1702 composite color monitor having a resolution of 320 by 200 pixels. Subjects controlled the cross-hair object with a standard digital, two-dimensional joystick. A firebutton mounted on the top of the joystick allowed subjects to fire at red objects. The same joystick was used for the location and identity probes.

## RESULTS

The results are reported under three major headings: reliability, criterion validity, and construct validity. Only Experiments 1 and 2 were used to evaluate reliability and criterion validity. For this evaluation, data from Experiment 1 was combined with data from the four Experiment 2 conditions in which the running memory task was omitted. Thus, data from a total of 32 subjects were used to evaluate reliability and criterion validity in the four experimental conditions common to both experiments. Results of the experimental manipulations themselves for the three individual experiments are reported under the heading of construct validity.

### Reliability

Between-session test-retest correlations for the SA metrics are displayed in Table 1. These are the correlations between sessions 2 and 3 for average location error, identity probe accuracy, identity probe response latency, and envelope sensitivity. (Correlations for the separate location errors by object identity were also obtained but did not differ appreciably from those for the global average reported here.) In addition, reliabilities for the measures of kill probability and avoidance failures are also shown. Pearson's product-moment correlations ($r$) were first computed individually for each of the four experimental conditions and then transformed to Fisher's $z$. The transformed correlations were then averaged across experimental conditions, and the obtained mean was back-transformed to Pearson's $r$. Statistical significance of the averaged correlations were tested using the procedure described by Dunlap, Silver, and Bittner (1936). Because the correlations were all expected to be positive, one-tailed tests were used.

Table 1. *Between-session test-retest correlations (Probability of Fisher's z; N = 32).*

| | Location Error | Identity | | Envelope |
|---|---|---|---|---|
| | | Accuracy | Latency | Sensitivity |
| Between-Session<br>(N=32, Exp 1 & 2) | .13<br>(ns) | .49<br>(.01) | .54<br>(.005) | .42<br>(.025) |

| | Kill Probability | Avoidance Failure |
|---|---|---|
| Between-Session<br>(N=32, Exp 1 & 2) | .48<br>(.01) | .26<br>(.10) |

Location errors were completely unreliable as is evident from the poor between-session correlations. In contrast, envelope sensitivity and both identity probe accuracy and latency produced test-retest correlations in the .4 to .5 range. Although statistically significant, these reliability coefficients may still be unacceptably low: they indicate that error accounts for half or more of the observed variability in the SA metrics (Murphy and Davidshofer, 1991).

One reason the reliability coefficients are important is because their square roots present theoretical upper limits to the validity coefficients that can be obtained (Allen and Yen, 1979). Specifically, the correlation between an SA metric and a another measure can not exceed the square root of the product of the two reliabilities. For example, the theoretical maximum correlation between identity probe latency and kill probability is the square root of .54 x .48, or .51.

### Criterion Validity

Correlations of the SA metrics with the criterion measure of kill probability are shown in the second row of Table 2 (the first row shows the theoretical upper limit of these correlations derived from the appropriate reliabilities--with the predicted sign of correlation added later). Note that correlations for envelope sensitivity are not shown because kill probability contributed to its computation; thus, the high correlation between these two variables is uninformative. Each correlation represents the average of eight individual correlations, one from each of the four conditions in each of the two test sessions. Again, Dunlap et al.'s (1986) procedure was used to test the statistical significance of the mean correlation. One-tailed tests were conducted because the sign of each correlation was predicted in advance of the experiment (e.g., if location error declined, then kill probability was expected to improve).

As can be seen, all the correlation coefficients were in the predicted direction as indicated by the sign of theoretical upper limit. Nevertheless, only the correlation between identity probe latency and kill probability achieved statistical significance: slower identity probe responses were associated with poorer kill probabilities. Note, however, that the obtained correlation is well below the theoretical limit and that the two measures had only about eight percent of their variance in common.

Table 2. *Theoretical upper limits (signs added) and observed correlations between SA metrics and kill probability (N = 32).*

| | Location | Identity | | Envelope |
|---|---|---|---|---|
| | Error | Accuracy | Latency | Sensitivity |
| Theoretical Limit | .25 | .48 | -.51 | - |
| Correlation with Kill Probability | .02<br>(ns) | .10<br>(ns) | -.29<br>(.05) | - |

## Construct Validity

Construct validity of the SA metrics was examined in three ways. First, the average correlation of each metric with avoidance failures was obtained. Second, the average correlations among the SA metrics themselves were computed. Third, the detailed results of the three experiments were analyzed.

### SA Metrics and Avoidance Failures

SA has been hypothesized to decline as task difficulty increases. Because avoidance failures were expected to increase with increasing difficulty, SA and avoidance should be related. Thus, location errors and identity probe latency should both have increased (indicating poorer SA) as avoidance failures increased. At the same time, envelope sensitivity and identity probe accuracy were expected to improve as avoidance failures declined. The correlations needed to test these predictions are shown in the second row of Table 3. The table also shows the correlation between kill probability and avoidance failures.

Table 3. *Theoretical upper limits (signs added) and observed correlations between SA metrics and avoidance failures (N = 32).*

| | Location | Identity | | Envelope | Kill |
|---|---|---|---|---|---|
| | Error | Accuracy | Latency | Sensitivity | Probability |
| Theoretical Limit | .18 | -.36 | .37 | -.33 | -.35 |
| Correlation with<br>Avoidance Failure | .10<br>(ns) | -.11<br>(ns) | .20<br>(.10) | -.39<br>(.025) | -.32<br>(.025) |

While all of the correlations were in the predicted direction, envelope sensitivity was the only SA metric to achieve a significant correlation with the avoidance failure measure. Although small, this correlation was as large as could be expected given the reliabilities of the measures. Table 3 also shows the correlation between kill probability and avoidance failures. Again, the correlation achieved statistical significance and approached its theoretical upper limit.

**Correlations Among SA Metrics**

If location error, identity probe performance, and envelope sensitivity all measure a common construct, then they should be correlated to some degree. On the other hand, the three measurement procedures could tap into different components of SA. If location error measures a spatial component while identity probes assess a verbal component, then those measures should be uncorrelated with each other. Because envelope sensitivity should reflect both spatial and verbal components, sensitivity should be correlated with both location and identity measures. Unfortunately, the extremely low reliability of location error makes the appearance of either pattern of correlations unlikely. Nevertheless, the correlations are reported for completeness.

Table 5 displays the correlations among the SA metrics. These obtained correlations may be compared to their theoretical upper limits (given their reliabilities) shown in Table 4.

Table 4. *Theoretical upper limits of the correlations among SA measures given their reliabilities (signs added to indicate expected direction of the correlations).*

| | Location Error | Identity Accuracy | Latency |
|---|---|---|---|
| Identity Accuracy | -.25 | | |
| Identity Latency | .26 | -.51 | |
| Envelope Sensitivity | -.23 | .45 | -.48 |

Table 5. *Obtained correlations among SA measures averaged across sessions and conditions (Probability of Fisher's Z, N = 32).*

| | Location Error | Identity Accuracy | Latency |
|---|---|---|---|
| Identity Accuracy | -.22<br>(.10) | | |
| Identity Latency | -.09<br>(ns) | -.28<br>(.05) | |
| Envelope Sensitivity | -.09<br>(ns) | .20<br>(ns) | -.31<br>(.025) |

Relative to its theoretical upper limit, the strongest correlation obtained was between location error and identity probe accuracy. On the other hand, the correlation of location error with identity probe latency was virtually zero. Envelope sensitivity appeared to be correlated with both identity probe accuracy and latency although only the correlation with latency was statistically significant.

**Experiment 1 Results**

Prior to analysis, all data for the four experimental conditions were averaged across the two replications within each session. The general procedure was to analyze the data in a 2 x 2 x 2 (session by enemy number by identity consistency) within-subjects multivariate analysis of variance (MANOVA) treating subjects as a random effect. If the MANOVA led to rejection of the null hypothesis, then univariate ANOVAs were carried out on the individual measures.

The dependent measures were analyzed in three separate MANOVAs corresponding to the three tasks. Session had no reliable effect in any of the analyses and so will not be discussed further.

*Combat task*. The primary measure of combat task performance was envelope sensitivity. Intuitively, this measure indicates subjects' ability to discriminate whether an enemy object was close enough to be destroyed. The second measure analyzed was the number of avoidance failures.

Table 6 displays the envelope sensitivity and avoidance failure data from the combat task. Only the number of enemy objects had a reliable effect on these two variables *Wilks' Lambda* = 0.11, $F(2,6) = 23.16$, $p < .002$. As enemy number increased from one to three, envelope sensitivity declined ($F(1,7) = 11.07$, MSe = 0.0006, $p < .02$) and avoidance failures increased ($F(1,7) = 21.15$, MSe = 6.7, $p < .003$). Neither the main effect of identity inconsistency nor the interaction with enemy number was reliable ($p$'s > .4).

Table 6. *Combat task data from Experiment 1.*

| | Envelope Sensitivity | Avoidance Failures |
|---|---|---|
| 1 Enemy | | |
| Identity | | |
| Consistent | .89 | 4.5 |
| Inconsistent | .88 | 3.5 |
| 3 Enemies | | |
| Identity | | |
| Consistent | .87 | 7.1 |
| Inconsistent | .87 | 6.9 |

*Location probe*. Table 7 displays the separate location errors for enemy, friendly, and neutral objects. Increasing the number of enemy objects had a reliable effect on these errors, *Wilks' Lambda* = 0.18, $F(3,5) = 7.47$, $p < .03$. The effect on enemy location error was the most dramatic with increasing enemy number leading to a large increase in error, $F(1,7) = 19.04$, MSe = 454.6, $p < .004$. A smaller increase observed in neutral location error was only marginally reliable, $F(1,7) = 6.26$, MSe = 253.1, $p < .05$. Friendly location error was unaffected ($p > .8$).

Table 7. *Location error (deg of visual angle) from Experiment 1.*

| | Object Identity | | |
|---|---|---|---|
| | Enemy | Friend | Neutral |
| 1 Enemy | | | |
| Identity | | | |
| Consistent | 2.4 | 6.5 | 6.2 |
| Inconsistent | 3.9 | 5.0 | 6.6 |
| 3 Enemies | | | |
| Identity | | | |
| Consistent | 5.8 | 5.2 | 6.8 |
| Inconsistent | 4.9 | 5.8 | 7.9 |

While identity consistency had no main effect on location errors ($p > .4$), it did reliably interact with enemy number, *Wilks' Lambda* = 0.12, $F(3,5) = 12.35$, $p < .01$). This interaction occurred only in enemy location error, $F(1,7) = 16.16$, MSe = 164.9, $p < .006$. The interaction was not reliable in either friendly or enemy location error ($p$'s > .14). From Table 7, it appears that enemy location error increased due to identity inconsistency when there was only one enemy object ($F(1,7) = 5.9$, $p < .10$) but not when there were three enemy objects ($p > .10$).

*Identity probe.* The reaction time and accuracy of responses in the identity probe are shown in Table 8. Both main effects and their interaction were reliable (Enemy Number: *Wilks' Lambda* = 0.06, $F(2,6) = 47.47$, $p < .0002$; Identity Consistency: *Wilks' Lambda* = 0.18, $F(2,6) = 13.83$, $p < .006$; Interaction: *Wilks' Lambda* = 0.34, $F(2,6) = 5.76$, $p < .05$). When enemy number increased, reaction time increased ($F(1,7) = 17.57$, MSe = 59919.6, $p < .005$) and accuracy decreased ($F(1,7) = 53.81$, MSe = 115.5, $p < .0002$). When object colors were inconsistent, reaction time was not affected ($p > .2$) but accuracy decreased, $F(1,7) = 24.19$, MSe = 139.0, $p < .002$. The number by identity consistency interaction was not reliable in the reaction time data ($p > .7$) but was in the accuracy data, $F(1,7) = 12.81$, MSe = 74.5, $p < .009$). Inspection of Table 8 suggests that the effect of identity inconsistency on response accuracy was much greater when there were three enemy objects rather than one. Statistical comparisons showed no reliable decrease in accuracy when there was only one enemy object ($p > .10$), but the decrease was reliable when there were three enemy objects, $F(1,7) = 25.9$, $p < .01$.

Table 8. *Identity probe data from Experiment 1.*

| | Reaction Time (ms) | Percent Correct |
|---|---|---|
| 1 Enemy | | |
| Identity | | |
| Consistent | 1241 | 93 |
| Inconsistent | 1159 | 86 |
| 3 Enemies | | |
| Identity | | |
| Consistent | 1483 | 81 |
| Inconsistent | 1430 | 59 |

**Experiment 2 Results**

The same statistical approach used in the first experiment was used here. Separate MANOVAS were carried out for the location probe, the identity probe, the combat task, and the memory tasks. In the case of the spatial memory task, where there was only one dependent variable, an ANOVA was used.

*Combat task.* Table 9 displays the combat task performance data for both the verbal and spatial memory subjects. The results from the first experiment were generally replicated with one exception. As in the previous experiment, the number of enemy objects had a reliable effect *Wilks' Lambda* = 0.09, $F(2,21) = 96.89$, $p < .0001$. As enemy number increased from one to three, envelope sensitivity declined ($F(1,22) = 94.33$, MSe = 0.002, $p < .0001$) and avoidance failures increased ($F(1,22) = 113.31$, MSe = 4.6, $p < .0001$). Also as in the previous experiment, the main effect of object-identity inconsistency was unreliable ($p > .16$). Unlike the previous experiment, however, the interaction between enemy number and identity inconsistency was reliable, *Wilks' Lambda* = 0.54, $F(2,21) = 9.00$, $p < .002$. This interaction did not appear in the avoidance failure data ($p > .17$) but only in envelope sensitivity, $F(1,22) = 12.98$, MSe = 0.0008, $p < .002$. Inspection of Table 9 suggests that identity inconsistency caused envelope sensitivity to decrease when there was one enemy object and to *increase* when there were three, but neither contrast was reliable when tested alone ($p$'s > .10). Thus, although the interaction is reliable, its contributing simple effects are not.

Table 9. *Combat task data from Experiment 2.*

| | 1 Enemy | | | | 3 Enemies | | | |
|---|---|---|---|---|---|---|---|---|
| | Envelope Sensitivity | | Avoidance Failures | | Envelope Sensitivity | | Avoidance Failures | |
| Identity | Vb | Sp | Vb | Sp | Vb | Sp | Vb | Sp |
| Consistent | | | | | | | | |
| 0 Items | .89 | .90 | 4.1 | 4.4 | .85 | .85 | 7.4 | 6.0 |
| 1 Item | .91 | .90 | 4.2 | 3.5 | .85 | .87 | 5.4 | 5.5 |
| 2 Items | .89 | .91 | 4.3 | 3.6 | .86 | .85 | 4.5 | 4.7 |
| Mean | .90 | .90 | 4.2 | 3.8 | .85 | .86 | 5.8 | 5.4 |
| Inconsistent | | | | | | | | |
| 0 Items | .89 | .90 | 4.0 | 3.1 | .86 | .86 | 6.3 | 5.3 |
| 1 Item | .88 | .90 | 3.7 | 3.0 | .85 | .87 | 6.2 | 5.0 |
| 2 Items | .89 | .90 | 3.4 | 1.9 | .87 | .87 | 5.0 | 4.7 |
| Mean | .89 | .90 | 3.7 | 2.7 | .86 | .87 | 5.8 | 5.0 |

*Ncte.* Vb and Sp identify the concurrent task as Verbal or Spatial.

Whether the memory span task was verbal or spatial had no effect on the combat task (all *p*'s > .30). Memory load did have a marginal effect on task performance, *Wilks' Lambda* = 0.76, *F*(4,86) = 3.09, *p* < .02. Although envelope sensitivity was not affected (*p* > .5), increasing memory load led to a *decrease* in SAM avoidance failures, *F*(2,44) = 6.59, *p* < .004, *Bonferroni p* < .007. This beneficial effect of memory load could have been due to the addition of the memory task (a load size of 0 versus 1, the "concurrency benefit"), the increase in load (1 versus 2, the "load effect"), or both. In order to evaluate these possibilities, the concurrency benefit and load effect contrasts were analyzed separately. Neither contrast appeared reliable when tested alone (*p*'s > .06). Thus, the significant overall effect must have been due to some combination of concurrency benefit and memory load.

*Location probe.* Table 10 displays the average location errors for enemy, friendly, and neutral objects. As in Experiment 1, increasing the number of enemy objects had a reliable effect on the location errors, *Wilks' Lambda* = 0.32, *F*(3,20) = 14.10, *p* < .0001. But unlike in the first experiment, the location errors for neutral objects was unaffected (*p* > .2). The increase in enemy location error was again dramatic, *F*(1,22) = 16.84, MSe = 1565.45, *p* < .0006. Friendly location error was again unaffected (*p* > .8).

Table 10. *Location error (deg of visual angle) from Experiment 2.*

| | 1 Enemy | | | | | | 3 Enemies | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Object Identity | | | | | | Object Identity | | | | | |
| | Enemy | | Friend | | Neutral | | Enemy | | Friend | | Neutral | |
| Identity | Vb | Sp | Vb | Sp | Vb | Sp | Vb | Sp | Vb | Sp | Vb | Sp |
| Consistent | | | | | | | | | | | | |
| 0 Items | 4.3 | 2.9 | 5.6 | 3.9 | 7.0 | 6.4 | 5.4 | 6.3 | 4.4 | 5.4 | 6.6 | 7.5 |
| 1 Item | 4.1 | 5.4 | 4.9 | 5.9 | 7.7 | 7.9 | 5.5 | 6.4 | 5.3 | 5.1 | 8.9 | 7.8 |
| 2 Items | 5.0 | 4.7 | 5.0 | 5.8 | 6.9 | 7.2 | 6.5 | 5.4 | 7.3 | 5.9 | 6.7 | 9.1 |
| Mean | 4.7 | 4.3 | 5.2 | 5.2 | 7.2 | 7.2 | 5.7 | 6.0 | 5.7 | 5.5 | 7.4 | 8.1 |
| Inconsistent | | | | | | | | | | | | |
| 0 Items | 5.0 | 4.6 | 4.8 | 4.5 | 7.1 | 7.6 | 5.2 | 6.0 | 5.8 | 5.1 | 6.9 | 7.6 |
| 1 Item | 4.7 | 4.4 | 9.8 | 4.8 | 7.7 | 7.2 | 5.7 | 6.5 | 6.0 | 5.0 | 7.2 | 9.1 |
| 2 Items | 5.2 | 5.5 | 6.2 | 6.3 | 7.7 | 7.6 | 6.3 | 5.9 | 5.3 | 5.9 | 7.0 | 7.5 |
| Mean | 5.0 | 4.8 | 6.9 | 5.2 | 7.5 | 7.5 | 5.7 | 6.1 | 5.7 | 5.3 | 7.0 | 8.1 |

Again identity inconsistency did not effect location errors ($p >$ .4). Further, the counter-intuitive interaction with enemy number that appeared in Experiment 1 was clearly unreliable here ($p > .5$).

Whether the memory task was verbal or spatial had no main effect on location errors ($p > .3$), and neither did memory load ($p > .16$). Further, memory type and load failed to interact with each other ($p >$ .7). However, memory type did interact with enemy object number, *Wilks' Lambda* = 0.67, $F(3,20) = 3.35$, $p < .04$. This interaction was non-significant in the location errors of both enemy and friendly objects (both $p$'s > .4) and was only marginally reliable in the errors for neutral objects, $F(1,22) = 3.71$, MSe = 1092.09, $p < .07$. Examination of Table 10 suggests that neutral object error was unaffected by enemy object number when the memory task was verbal; but when the memory task was spatial, increased enemy object number caused neutral object error to increase.

No other main or interaction effects were reliable in the location probe data (all $p$'s > .2) with one exception. The four-way interaction between memory type, memory load, enemy object number, and identity inconsistency was reliable in the MANOVA, *Wilks' Lambda* = 0.72, $F(6,84) = 2.49$, $p < .03$. Examination of the three univariate ANOVA's revealed that this interaction was unreliable in enemy object error ($p > .6$) and only marginally reliable in friendly object error, $F(2,44) = 2.44$, MSe = 2566.14, $p < .10$, and in neutral object error, $F(2,44) = 2.56$, MSe = 1279.73, $p < .09$. Because the reliability of these interactions is questionable, and because each interaction involves several possible contrasts among the 24 different means, further analysis was not deemed worthwhile.

*Identity probe.* The major effects of the first experiment on the identity probe were also replicated. The reaction time and accuracy of identity probe responses are shown in Table 11. Main effects of both enemy object number and identity inconsistency were both reliable as was their interaction (Enemy Number: *Wilks' Lambda* = 0.13, $F(2,21) = 72.02$, $p < .0001$; Identity Consistency: *Wilks' Lambda* = 0.37, $F(2,21) = 17.52$, $p < .0001$; Interaction: *Wilks' Lambda* = 0.54, $F(2,21) = 9.01$, $p < .002$). When enemy number increased, reaction time increased ($F(1,22) = 8.54$, MSe = 311894.6, $p < .008$) and accuracy decreased ($F(1,21) = 140.89$, MSe = 333.2, $p < .0001$). When colors were inconsistent, reaction time was again unaffected ($p > .6$) but accuracy again decreased, $F(1,22) = 36.51$, MSe = 604.8, $p < .0001$. Again, the number by identity consistency interaction was not reliable in the reaction time data ($p > .8$) but was in the accuracy data, $F(1,22) = 17.84$, MSe = 327.1, $p < .0003$). Examination of Table 11 reveals the same pattern as before: the effect of identity inconsistency on response accuracy was much greater when there were three enemies rather than one. Statistical comparisons showed no reliable decrease in accuracy when there was only one enemy ($p > .10$); the decrease was reliable when there were three enemies, $F(1,22) = 6.45$, $p < .03$.

Table 11. *Identity probe data from Experiment 2.*

| | 1 Enemy | | | | 3 Enemies | | | |
|---|---|---|---|---|---|---|---|---|
| | Reaction Time (ms) | | Percent Correct | | Reaction Time (ms) | | Percent Correct | |
| Identity | Vb | Sp | Vb | Sp | Vb | Sp | Vb | Sp |
| Consistent | | | | | | | | |
| 0 Items | 1365 | 1321 | 83 | 86 | 1525 | 1332 | 71 | 75 |
| 1 Item | 1211 | 1281 | 83 | 81 | 1486 | 1458 | 74 | 72 |
| 2 Items | 1368 | 1293 | 82 | 88 | 1443 | 1447 | 72 | 68 |
| Mean | 1314 | 1298 | 83 | 85 | 1485 | 1412 | 72 | 72 |
| Inconsistent | | | | | | | | |
| 0 Items | 1225 | 1306 | 83 | 80 | 1537 | 1372 | 55 | 61 |
| 1 Item | 1292 | 1323 | 79 | 77 | 1491 | 1443 | 49 | 55 |
| 2 Items | 1448 | 1399 | 74 | 74 | 1531 | 1398 | 47 | 53 |
| Mean | 1322 | 1343 | 78 | 77 | 1520 | 1404 | 50 | 56 |

No other main or interaction effects were reliable ($p$'s > .1). Thus, neither memory type nor load had any reliable effects on the identity probe.

*Memory tasks.* The data for the verbal and spatial memory tasks are shown in Table 12. Both the verbal and spatial memory tasks were affected by combat task intensity; Verbal: *Wilks' Lambda* = 0.41, $F(2,10) = 7.25$, $p < .02$; Spatial: $F(1,11) = 6.26$, *MSe* = 379.0, $p < .03$. Table 12 shows that verbal memory accuracy declined as enemy number increased, *F*(1,11) = 11.37, MSe = 697.1, *p* < .007, although reaction time was unaffected, *p* > .5. From the table, it is apparent that spatial memory error increased. Neither memory task was affected by identity inconsistency; further, identity inconsistency did not interact with any other variables in either task (all *p*'s > .4).

Table 12. *Memory task data from Experiment 2.*

| | 1 Enemy | | | 2 Enemies | | |
|---|---|---|---|---|---|---|
| | Verbal | | Spatial | Verbal | | Spatial |
| | RT | Percent | Error | RT | Percent | Error |
| Identity | (ms) | Correct | (Deg) | (ms) | Correct | (Deg) |
| Consistent | | | | | | |
| 1 Item | 1226 | 54 | 6.0 | 1193 | 44 | 7.0 |
| 2 Items | 1889 | 56 | 7.2 | 1681 | 42 | 8.2 |
| Mean | 1558 | 55 | 6.6 | 1437 | 43 | 7.6 |
| Inconsistent | | | | | | |
| 1 Item | 1167 | 55 | 6.0 | 1285 | 43 | 6.5 |
| 2 Items | 1959 | 55 | 7.1 | 1913 | 38 | 8.1 |
| Mean | 1563 | 55 | 6.5 | 1599 | 41 | 7.3 |

The only other reliable effect in either task was that of memory load Verbal: *Wilks' Lambda* = 0.26, *F*(2,10) = 13.87, *p* < .002; Spatial: *F*(1,11) = 13.15, MSe = 815.9, *p* < .004. Increasing memory load from one to two stimuli caused spatial error to increase. In the verbal task, reaction time increased, *F*(1,11) = 28.9, MSe = 685225.7, *p* < .0002; accuracy was unaffected (*p* > .7).

**Experiment 3 Results**

Prior to the main statistical analysis, the 2 x 3 x 3 (Session by Verbal Load by Spatial Load) data matrix was reduced to a 2 x 5 (Session by Combined Load) matrix. The five combined loads were 0, 1, 2, 3, and 4 items to be remembered. From the original three-way matrix, there was only one condition that contributed to the 0-item cell in the new two-way matrix, two conditions that contributed to the 1-item cell, three conditions that contributed to the 2-item cell, two conditions that contributed to the 3-item cell, and one condition that contributed to the 4-item cell. Therefore, in order to obtain the same number of observations in each cell, each subject's data for each cell were averaged across the contributing conditions.

Once the new two-way data matrix was created, the same MANOVA strategy used in the previous experiments was used. If the MANOVA revealed an effect on a task, then ANOVAs were applied to the individual dependent measures. If the ANOVA identified a significant effect of combined memory load, the effect was analyzed further using orthogonal polynomial contrasts in order to determine the shape of the effect. These polynomial contrasts were computed individually for each subject so that each contrast could be tested against its own error term.

*Combat task; location and identity probes.* Tables 13, 14, and 15 display the data from the combat, location, and identity probes, respectively. Combined memory load had no reliable effect on the combat task ($p > .1$) nor on the location or identity probes (all $p$'s > .4). However, visual inspection of the location error data for enemy objects suggested that error may have increased as memory load increased from 0 to 3 items, and then decreased when load increased to 4 items. In order to determine whether this trend might be real in spite of the insignificant global MANOVA, a Bonferroni $F$ was used to evaluate the univariate ANOVA test of the effect of memory load on enemy object error. The trend was found to be unreliable ($p > .18$).

Table 13. *Combat task data from Experiment 3.*

| | Envelope Sensitivity | Avoidance Failures |
|---|---|---|
| 0 Items | .85 | 4.2 |
| 1 Item | .83 | 5.8 |
| 2 Items | .82 | 5.4 |
| 3 Items | .82 | 6.9 |
| 4 Items | .82 | 5.7 |

Table 14. *Location error (deg of visual angle) from Experiment 3.*

| | Object Identity | | |
|---|---|---|---|
| | Enemy | Friend | Neutral |
| 0 Items | 4.9 | 5.9 | 6.6 |
| 1 Item | 5.4 | 5.9 | 7.2 |
| 2 Items | 5.9 | 6.6 | 6.9 |
| 3 Items | 6.5 | 6.3 | 7.4 |
| 4 Items | 5.1 | 6.6 | 6.8 |

Table 15. *Identity probe data from Experiment 3.*

| | Reaction Time (ms) | Percent Correct |
|---|---|---|
| 0 Items | 2245 | 52 |
| 1 Item | 2661 | 48 |
| 2 Items | 2370 | 48 |
| 3 Items | 2461 | 51 |
| 4 Items | 2352 | 51 |

*Memory tasks.* Data from the memory tasks are displayed in Table 16. Both the verbal and spatial memory tasks were affected by combined memory load (verbal: *Wilks' Lambda* = 0.27, *F*(6,46) = 7.04, *p* < .0001; spatial: *F*(3,24) = 4.48, MSe = 1678.9, *p* < .02). In the verbal task, both reaction time and accuracy were affected (reaction time: *F*(3,24) = 12.09, MSe = 801343.4, *p* < .0001; accuracy: *F*(3,24) = 4.36, MSe = 389.9, *p* < .02). Table 16 suggests that verbal reaction time increased as the combined span increased from 1 to 4 items. This conclusion is borne out by the polynomial contrasts: the linear contrast was reliable, *F*(1,8) = 46.63, MSe = 118222.1, *p* < .0001, but neither the quadratic nor the third degree contrasts were significant (*p*'s > .2). At the same time, verbal accuracy appeared to decrease when memory load increased from 1 to 2 but was unaffected thereafter. Thus, neither the linear nor third degree contrasts in accuracy were significant (*p* > .09, *p* > .2, respectively), but the quadratic contrast was reliable, *F*(1,8) = 6.09, MSe = 1028.6, *p* < .04. Finally, error in the spatial task also appears to have increased regularly as memory load increased from 1 to 3 items (and was then unaffected by a further increase to 4 items). As a result, the linear contrast was significant, *F*(1,8) = 8.31, MSe = 1073.4, *p* < .03, but neither the quadratic nor the third degree contrasts were reliable (*p*'s > .3).

Table 16. *Memory task data from Experiment 3.*

| | Verbal Memory | | Spatial Memory |
|---|---|---|---|
| | Reaction Time (ms) | Percent Correct | Error (Deg of vis angle) |
| 1 Item | 1321 | 60 | 5.9 |
| 2 Items | 1536 | 46 | 6.7 |
| 3 Items | 1598 | 37 | 7.8 |
| 4 Items | 1883 | 49 | 7.7 |

## DISCUSSION

These experiments examined the psychometric characteristics of both explicit and implicit measures of SA. The focus was on their reliability, criterion validity, and construct validity. Each of these is discussed in turn.

### Reliability

None of the evaluated metrics exhibited a high degree of reliability. The most reliable of the measures, identity probe latency, produced a test-retest reliability coefficient of less than .60--which most researchers would consider to be unacceptably low (Murphy and Davidshofer, 1991). Location probes proved even more unreliable, yielding a reliability coefficient that was effectively zero. These results raise at least two questions: (1) why did the SA metrics prove unreliable, and (2) why were location probes so much less reliable than other measures?

In answer to the first question, Allen and Yen (1979) suggested three possibilities that should be considered. First, some subjects may have benefited more than others from practice across sessions. Second, some subjects' attitudes toward the experiment may have changed in between sessions. For example, a subject who was cooperative during one test session may have been uncooperative in the other. Third, some subjects may have used the period between sessions to devise a more effective strategy for coping with task demands while others may have become ill or obtained too little sleep. Any one of these three possibilities is sufficient to change the rank order of subjects on a given measure from one session to another; thus, any one or all could account for the low test-retest correlation.

In the present experiments, test-retest correlations were taken between sessions two and three in the hope that the major contaminating effects would occur prior to session two. Statistical comparison of session means showed that no systematic changes occurred in any of the measures between sessions two and three. Nevertheless, if some subjects improved across sessions while others grew worse, the rank ordering of subjects would have changed even though the global session means remained the same. Thus, the absence of a global session effect does not rule out the possibility of idiosyncratic session effects for individual subjects. As a result, it seems likely that the obtained reliability coefficients underestimate the actual reliabilities. Future research could overcome this difficulty by using internal consistency rather than test-retest correlations to estimate reliability. For example, subjects could perform under several (say, twelve) replications of the same condition in a given session, and one could then correlate average scores for even numbered replications with those for odd numbered replications. Spearman-Brown estimates of reliability could then be computed from the obtained correlations (Allen and Yen, 1979).

While idiosyncratic session effects may explain the poor test-retest reliability of the SA metrics overall, they probably cannot account for the much poorer, near zero reliability of location probes. At present, the reasons for location error unreliability are unknown; nevertheless, at least two plausible hypotheses may be offered. First, subjects may have encoded object locations more poorly than object identities. Because objects retained their identities for much longer periods of time than they remained at specific locations, encoded object-location associations would be weaker and subject to more proactive interference from previous associations than would object identities. Consequently, a good deal more error would be expected in responses to location rather than identity probes, thereby producing the much lower reliabilities. Unfortunately, this account--if correct--suggests that there may be little that can be done to improve location probe reliability relative to the other SA metrics

On the other hand, subjects may have coped with the difficulty of attended to exact pixel locations by adopting a different strategy altogther. For example, subjects may have attended only to general regions on the screen within which objects were located. If so, then subjects may have been content simply to place objects within the correct region without regard to its precise location inside that region. Then the smallest unit of error that was psychologically meaningful may have been much larger than the units (pixels) in which error was measured. In other words, subjects may have been responding to the location probes with less precision than the measurement process assumed. For example, if the smallest psychologically significant region was 15 pixels square, then subjects would not have distinguished among any errors smaller than 15 pixels. Determining reliability with individual pixels as units would then produce the near zero reliabilities observed here even though, in terms of psychological units, location error may in fact have been highly consistent across sessions. In order to evaluate this possibility, future research could seek to identify the size of the psychological unit and then rescale location error in terms of this unit. An alternative to rescaling would be to constrain subjects to only a few discrete locations; then the subjects' task would be to choose which location was correct.

**Criterion Validity**

In light of their limited reliability, the criterion validity of both identity and location probe measures may be better than the obtained correlations suggest. These are discussed in turn. As noted, envelope sensitivity was computed from kill probability combined with false alarms (firing when no enemy was present); thus, a meaningful assessment of criterion validity was not possible.

Not surprisingly, knowing which objects are friend, foe, and neutral appears to be a fairly good predictor of kill probability. Although the validity coefficients obtained for identity probe accuracy and latency are small, they are not at all uncommon in applied research (Murphy and Davidshofer, 1991). These low coefficients are no doubt due in part to the low reliability of the accuracy and latency measures. In order to estimate what the validity coefficients would have been if all the measures had been perfectly reliable, Spearman's (1904) correction for attenuation was computed (coefficients in the second row of Table 2 divided by the unsigned coefficients in the first row). This correction, which almost certainly overestimates a measure's potential criterion validity, yielded coefficients of .21 and -.57, respectively. Identity probe latency thus may be both an acceptable predictor of kill probability and a better predictor than identification accuracy. This superiority of latency compared to accuracy is not too surprising given that subjects had unlimited time to respond to the identity probes. Subjects could compensate for weakly encoded object identities by taking more time to retrieve the correct response. Thus, "weak" encodings--which would lower kill probability--would be more likely to influence latency than accuracy.

What may seem strange at first is the apparent independence of kill probability from average location error. Even after correction for attenuation, the validity coefficient was still near zero (.08). A likely explanation is that the criterion validity of average location error is attenuated by the inclusion of location errors for friendly and neutral objects--both of which were irrelevant to kill probability. Enemy location error--included as a component of the composite average--was expected to be even less reliable than average location error, and it was (*r* = .06). Nevertheless, the obtained correlation with kill probability was stronger, *r* = -.11. When corrected for attenuation due to unreliability, this correlation increased to -.65. Thus, it appears that a more reliable measure of enemy location error could be a good predictor of kill probability.

Interestingly, the best predictor of the kill probability criterion turned out not to be an SA metric at all, but rather a workload metric: specifically, performance on the embedded avoidance task. The correlation between avoidance failures and kill probability was higher than that observed for identification latency, and reached -.91 when corrected for attenuation due to unreliability. As a purely practical matter, this result might suggest that mental workload rather than situation awareness may be more useful in predicting at least some performance criteria.

**Construct Validity**

In the present experiments, construct validity was evaluated both between and within experimental conditions. Comparisons of measures between conditions examined whether the various measures reflected the hypothesized structure of processing resources thought to underlie SA. Within individual conditions, the relationships among different metrics provided additional insight into what memory probes and envelope sensitivity might actually measure.

*Between conditions.* The predictions used to assess construct validity were generated by the hypotheses that (1) SA in the combat task consisted of object locations and object identities, and (2) object locations would be maintained in a spatial memory while identities would be maintained in a verbal memory. Following Wickens (1984) multiple resource hypothesis which holds that spatial processes do not interfere with verbal processes, it was expected that increasing the difficulty of maintaining object identities would not affect location error--only the identity probe measures would be influenced. This prediction was confirmed in both Experiments 1 and 2 where identity inconsistency was detrimental to identity probe performance (responses became both slower and less accurate) but had no effect on location error. In addition, increasing the intensity of the combat task was predicted to overtax the resources available to the component processes underlying both location and identity awareness. Again, this prediction was confirmed in both experiments 1 and 2 where (a) enemy location error increased unaer greater combat intensity and (b) the detrimental effect of identity inconsistency on identity probe performance was exacerbated.

While the insensitivity of location error to identity inconsistency supports the separate resources hypothesis, an alternative explanation is that the apparent insensitivity was due to the unreliability of location error. Perhaps a more reliable measure of spatial awareness would have revealed an effect of identity inconsistency. Although this possibility cannot be ruled out by the present experiments, the avoidance failure data indicate that the separate resources hypothesis still may be correct. Specifically, avoidance failures became more likely when enemy number increased, but were unaffected by identity inconsistency. Given separate spatial and verbal resources, this result is expected if identity processing draws upon verbal resources while the avoidance task demands spatial resources. Less clear are the implications of the envelope sensitivity data. Logically, envelope sensitivity should have been influenced by both enemy number and identity inconsistency, showing the same interaction found in identity probe accuracy data; yet sensitivity was affected only by enemy number and not by identity inconsistency.

If location and identity awareness draw upon separate processing resources, these resources may not be characterized simply as spatial versus verbal. In Experiments 2 and 3, neither spatial nor verbal memory load had any apparent effect on the location or identity probe measures. Further, Experiment 3 raises doubts that spatial and verbal resources are really even separate because spatial and verbal spans clearly combined to overload working memory, a result that should be impossible if spatial and verbal resources are separate. This is not to say that separate resources do not exist, only that separate spatial and verbal resources appear unlikely. Separate resources of some as yet unknown characterization could still account for the insensitivity of location error to identity inconsistency. On the other hand, perhaps a non-resource explanation for the present results might be appropriate, such as Navon's outcome conflict theory (Navon, 1984, 1985, 1990; Navon and Miller, 1987; see also Fracker and Wickens, 1989; Hirst and Kalmar, 1987).

*Within conditions.* The picture presented within experimental conditions is not any clearer. The expected pattern of correlations between the SA metrics and avoidance failures did not occur. Because avoidance failures were thought to reflect overloaded spatial resources, a high correlation with location error was expected while correlations with the identity probe measures should have been low. This pattern was not found: identity probe latency showed a stronger correlation with avoidance failure than did location error. When corrected for attenuation due to unreliability, location error and identity probe latency were equally well correlated with avoidance failure (corrected *r*'s were about .5). Interestingly, envelope sensitivity was the SA measure most strongly correlated with avoidance failure--perhaps because both measures were collected during the combat task rather than during different phases of the experiment (i.e., combat task versus the probe freezes). Contrary to the hypothesis originally motivating the present research, these correlations suggest that a common processing resource may have contributed to performance on all of the SA measures.

Perhaps a common resource also explains the pattern of correlations expected among the SA metrics themselves. Identity probe accuracy was predictably correlated with latency but also with location error. At the same time, latency was correlated with envelope sensitivity. Thus, location probes, identity probes, and envelope sensitivity may all reflect some common process in addition (perhaps) to something unique to each metric. Whether this common process should be characterized as mental workload or as something else remains for future research to clarify.

**Conclusions**

Evidently, there is still much work to be done in order to develop SA metrics that are both reliable and valid. Changes in how the measures are implemented may improve reliability and should be pursued. Indeed, a final assessment of validity cannot be achieved until the measures have been made--or shown--to be reliable. In terms of criterion validity, tne present data suggest that SA metrics may be less useful than measures of workload, a sobering if not altogether surprising result. In terms of construct validity, the present data may be taken as evidence against either the validity of the measures or the usefulness of current multiple resource conceptions. Deciding between these two possibilities will have to await future research.

Finally, while the specific measures evaluated in this study may sample components of situation awareness, they probably do not exhaust the construct (Fracker, 1988; Sarter and Woods, 1991). New measures may well still be needed in order to capture other aspects of operator SA, especially the operator's comprehension of the "big picture," or how the elements of a situation fit together and form a comprehensive whole. One suspects that developing such global measures will not be easy, and that establishing their reliability and validity will be a major challenge.

## REFERENCES

Allen, M. J., and Yen, W. M. (1979). *Introduction to measurement theory.* Monterey, CA: Brooks-Cole.

Baddeley, A. D. (1986). *Working memory.* Oxford: Clarendon Press.

Carr, T. H. (1986). Perceiving visual language. In K. R. Boff, L. Kauffman, and J. P. Thomas (Eds.), *Handbook o. perception and human performance, Volume II: Cognitive Processes and Performance* (29:1-92). New York: Wiley.

Clausewitz, C. V. (1836/1984). *On war.* Princeton, NJ: Princeton University Press.

Craig, A. (1979). Nonparametric measures of sensory efficiency for sustained monitoring tasks. *Human Factors, 21,* 69-78.

Downing, C. J. (1988). Expectancy and visual-spatial attention: Effects of perceptual quality. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 188-202.

Dunlap, W. P., Silver, N. C., and Bittner, A. C. (1986). Estimating reliability with small samples: Increased precision with averaged correlations. *Human Factors, 28,* 685-690.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd annual meeting.* Santa Monica, CA: Human Factors Society.

Endsley, M. R. (1989). A methodology for the objective measurement of pilot situation awareness. In *Proceedings of the NATO AGARD Conference on Situational Awareness in Aerospace Operations* (AGARD-CP-478). Springfield, VA: National Technical Information Service.

Eriksen, C. W., and Yeh, Y. (1985). Attention allocation in the visual field. *Journal of Experimental Psychology: Human Perception and Performance, 11,* 583-597.

Eubanks, J. L., and Killeen, P. R. (1983). An application of signal detection theory to air combat training. *Human Factors,* 25, 449-456.

Fracker, M. L., & Wickens, C. D. (1989). Resources, confusions, and compatibility in dual-axis tracking: Displays, controls, and dynamics. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 80-96.

Friedman, A., & Polson, M. C. (1981). The hemispheres as independent resources: Limited capacity processing and cerebral specialization.

*Journal of Experimental Psychology: Human Perception and Performance, 7,* 1031-1058.

Gopher, D., and Donchin, D. (1986). Workload: An examination of the concept. In K. R. Boff, L. Kauffman, and J. P. Thomas (Eds.), *Handbook of perception and human performance, Volume II: Cognitive processes and performance* (41:1-49). New York: Wiley.

Gulliksen, H. (1950). *Theory of mental tests.* New York: Wiley.

Hawkins, H. L., Hillyard, S. A, Luck, S. J., Mouloua, M., Downing, C. J., and Woodward, D. P. (1990). Visual attention modulates signal detectability. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 802-811.

Hirst, W., & Kalmar, D. (1987). Characterizing attentional resources. *Journal of Experimental Psychology: General, 116,* 68-81.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Kantowitz, B. H. (1985). Channels and stages in human information processing: A limited analysis of theory and methodology. *Journal of Mathematical Psychology,* 29, 135-174.

Kibbe, M. P. (1988). Information transfer from intelligent EW displays. In *Proceedings of the Human Factors Society 32nd annual meeting.* Santa Monica, CA: Human Factors Society.

Loftus, E. F. (1979). *Eyewitness testimony.* Cambridge, MA: Harvard University Press.

Loftus, E. F., and Loftus, G. R. (1980). On the permanence on stored information in the human brain. *American Psychologist, 35,* 409-420.

Lord, F. M., and Novick, M. R. (1968). *Statistical theories of mental test scores.* Menlo Park, CA: Addison-Wesley.

Macmillan, N. A., & Creelman, C. D. (1990). Response bias: Characteristics of detection theory, threshold theory, and "non-parametric" indexes. *Psychological Bulletin, 107,* 401- 413.

Marshak, W. P., Kuperman, G., Ramsey, E. G., and Wilson, D. (1987). Situation awareness in map displays. In *Proceedings of the Human Factors Society 31st annual meeting* (pp. 533-538). Santa Monica, CA: Human Factors Society.

Moray, N., (Ed.) (1979). *Mental workload.* New York: Plenum.

Moray, N. (1988). Mental workload since 1979. *International Review of Ergonomics, 2,* 123-150.

Murphy, K. R., and Davidshofer, C. O. (1991). *Psychological testing: Principles and applications.* Englewood Cliffs, NJ: Prentice Hall.

Navon, D. (1984). Resources--A theoretical soupstone? *Psychological Review, 91,* 216-234.

Navon, D. (1985). *Do people allocate limited resources among concurrent activities?* (Technical Report No. IPDM 26, April). Haifa, Israel: University of Haifa, Laboratory for Information Processing and Decision Making.

Navon, D. (1990). Exploring two methods for estimating performance tradeoff. *Bulletin of the Psychonomic Society, 28,* 155-157.

Navon, D., and Miller, J. (1987). Role of outcome conflict in dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 435-448.

Norman, D. A., and Bobrow, D. G. (1975). On data-limited and resource- limited processes. *Cognitive Psychology, 7,* 44-64.

O'Donnell, R. D., and Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kauffman, and J. P. Thomas (Eds.), *Handbook of perception and human performance, Volume II: Cognitive Processes and Performance* (42:1-49). New York: Wiley.

Ogden, G., Levine, J., and Eisner, E. (1979). Measurement of workload by secondary tasks. *Human Factors, 21,* 529-548.

Pollack, I., & Norman, D. A. (1964). A nonparametric analysis of recognition experiments. *Psychonomic Science,* 1, 125-126.

Roediger, H. L. (1990). Implicit memory: Retention without remembering. *American Psychologist, 45,* 1043-1056.

Sarter, N. B., and Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon. *The International Journal of Aviation Psychology, 1,* 45-57.

Schneider, W., & Detweiler, M. (1988). The role of practice in dual-task performance: Toward workload modeling in a connectionist/control architecture. *Human Factors, 30,* 539- 566.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15,* 72-101.

Venturino, M., Hamilton, W. L., and Dvorchak, S. R. (1989).

Performance- based measures of merit for tactical situation awareness. In *Proceedings of the NATO AGARD conference on situational awareness in aerospace operations* (AGARD-CP-478). Springfield, VA: National Technical Information Service.

Venturino, M., and Kunze, R. J. (1989). Spatial awareness with a helmet mounted display. In *Proceedings of the Human Factors Society 33rd annual meeting* (pp. 1388-1391). Santa Monica, CA: Human Factors Society.

Wells, M. J., Venturino, M., and Osgood, R. K. (1988). Using target replacement performance to measure spatial awareness in a helmet-mounted display. In *Proceedings of the Human Factors Society 32nd annual meeting* (pp. 1429-1433). Santa Monica, CA: Human Factors Society.

Whitaker, L. A., and Klein, G. A. (1988). Situation awareness in the virtual world. In *Proceedings of the eleventh symposium on psychology in the Department of Defense* (USAFA-TR-88-1, pp. 321-325). United States Air Force Academy.

Wickens, C. D. (1980). The structure of attentional resources. In R. Nickerson & R. Pew (Eds.), *Attention and performance VIII* (pp. 239- 258). Hillsdale, NJ: Erlbaum.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63-102). New York: Academic Press.

Wickens, C. D. (in press). Processing resources and attention. In D. Damos (Ed.), *Multiple task performance*. London: Taylor & Francis.

Wickens, C. D., & Liu, Y. (1988). Codes and modalities in multiple resources: A success and a qualification. *Human Factors, 30*, 599-616.

Wickens, C. D., Sandry, D. L., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output. *Human Factors, 25*, 227- 248.

Wierwille, W. (1979). Physiological measures of aircrew mental workload. *Human Factors, 21*, 575-593.

Williges, R., and Wierwille, W. (1979). Behavioral measures of aircrew mental workload. *Human Factors, 21*, 549-574.

U.S. Government Printing Office [illegible]